Rare-to-Frequent: Unlocking Compositional Generation Power of Diffusion Models on Rare Concepts with LLM Guidance

Dongmin Park[1], Sebin Kim[2], Taehong Moon[1], Minkyu Kim[1], Kangwook Lee[1,3], Jaewoong Cho[1]

[1]KRAFTON [2] [3]WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

Project Page Dongmin

Rare Concept Composition for Visual Generation

- Generating images from prompts with **rare concept mixtures**
  - Object with unusual attributes
  - e.g., *a hairy frog, a star-shaped apple, a trumpet-like gun, …*
- Essential for real creators designing images *never seen* before

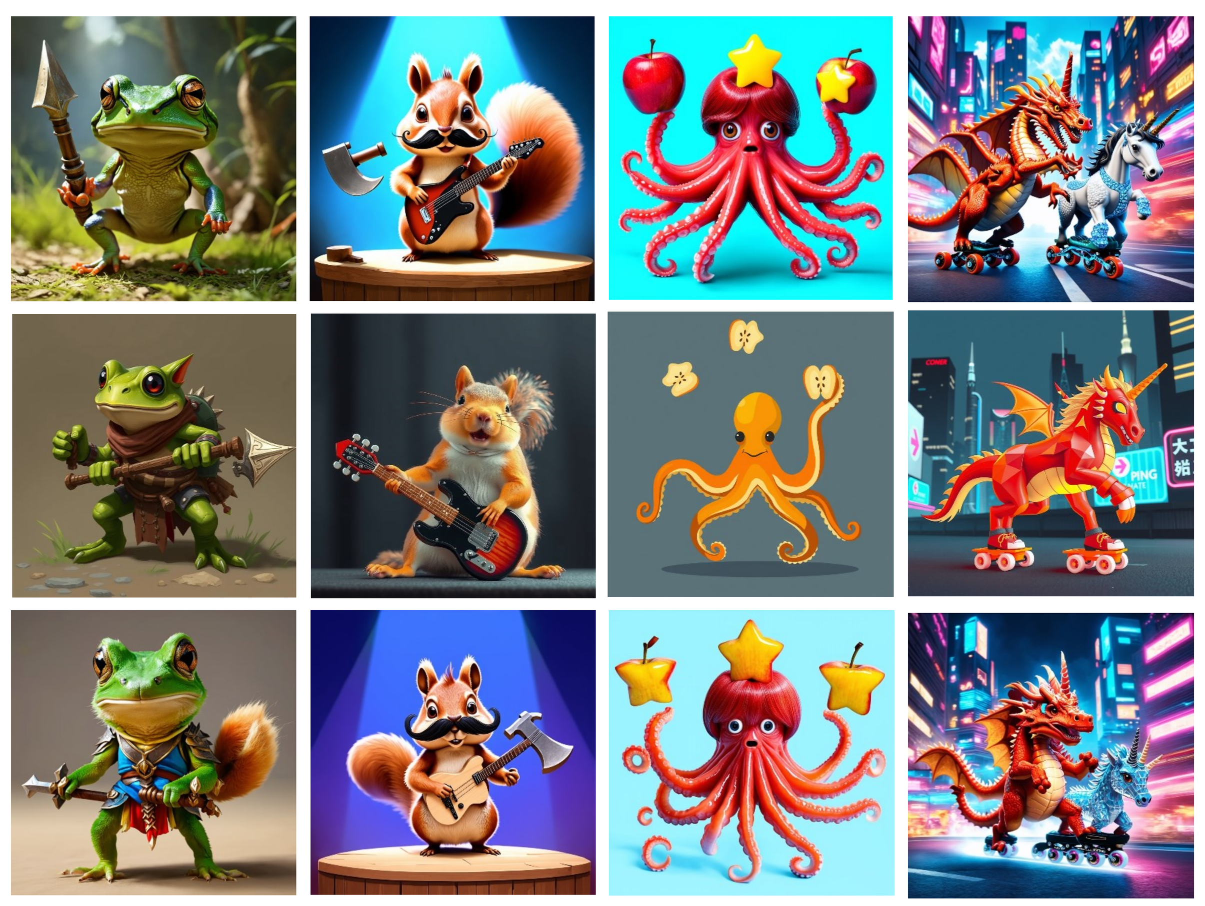
T2I Model's Performance?

- SOTA T2I models like **SD3 & FLUX often fail to generate**
- This may significantly degrade the user satisfaction

SD3

FLUX

R2F (OURS)

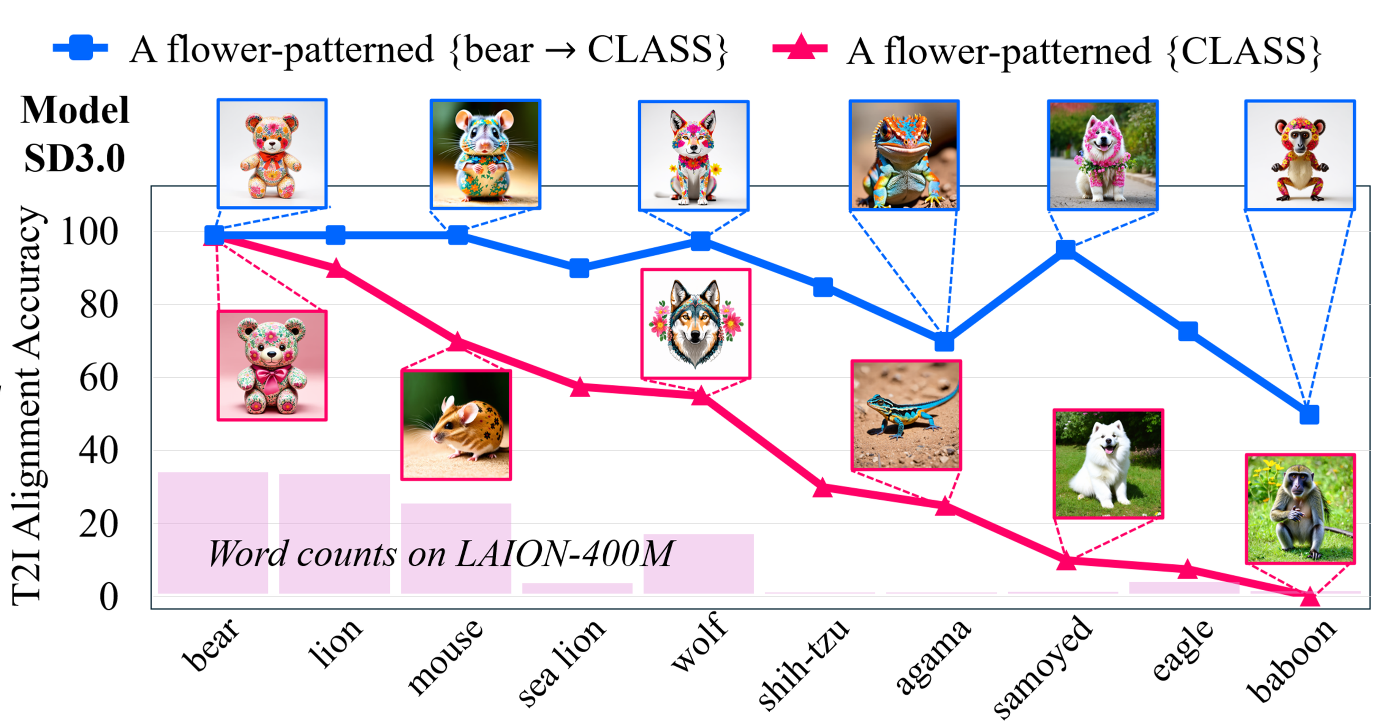


Hairy frog warrior | *Mustachioed squirrel with ax-shaped guitar* | *Wigged octopus juggling three star-shaped apples* | *Red dragon and diamond unicorn rollerblading*

Exposing Frequent Concept Improves Composition

- Obs)

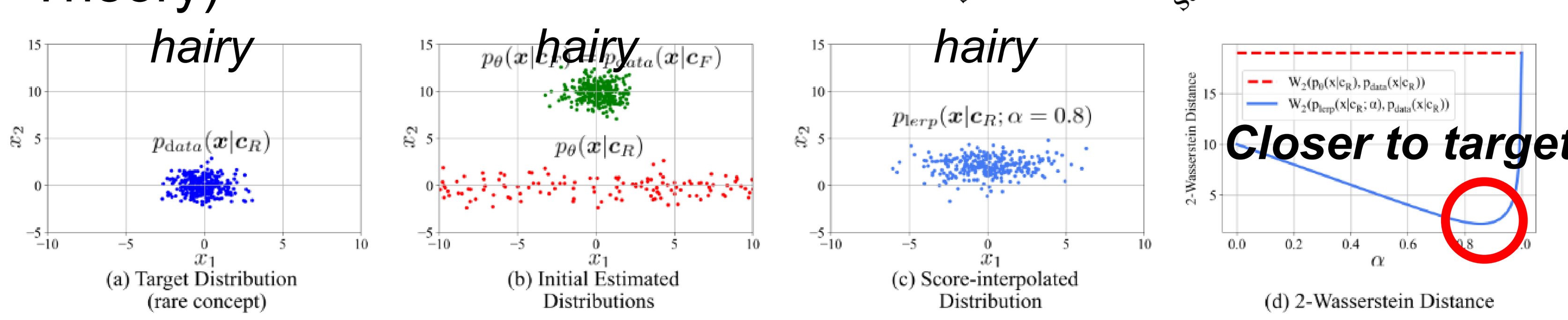
Freq concept in early-steps only!



- Theory)

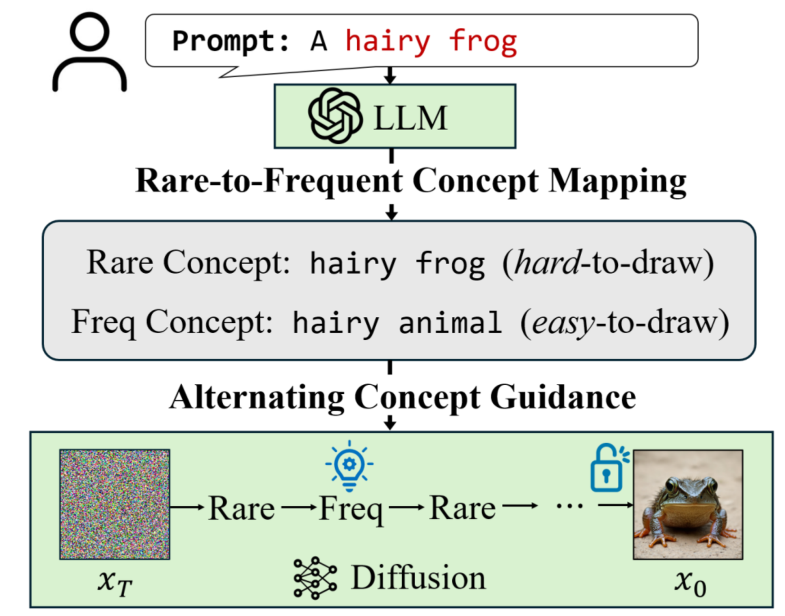
dog

frog

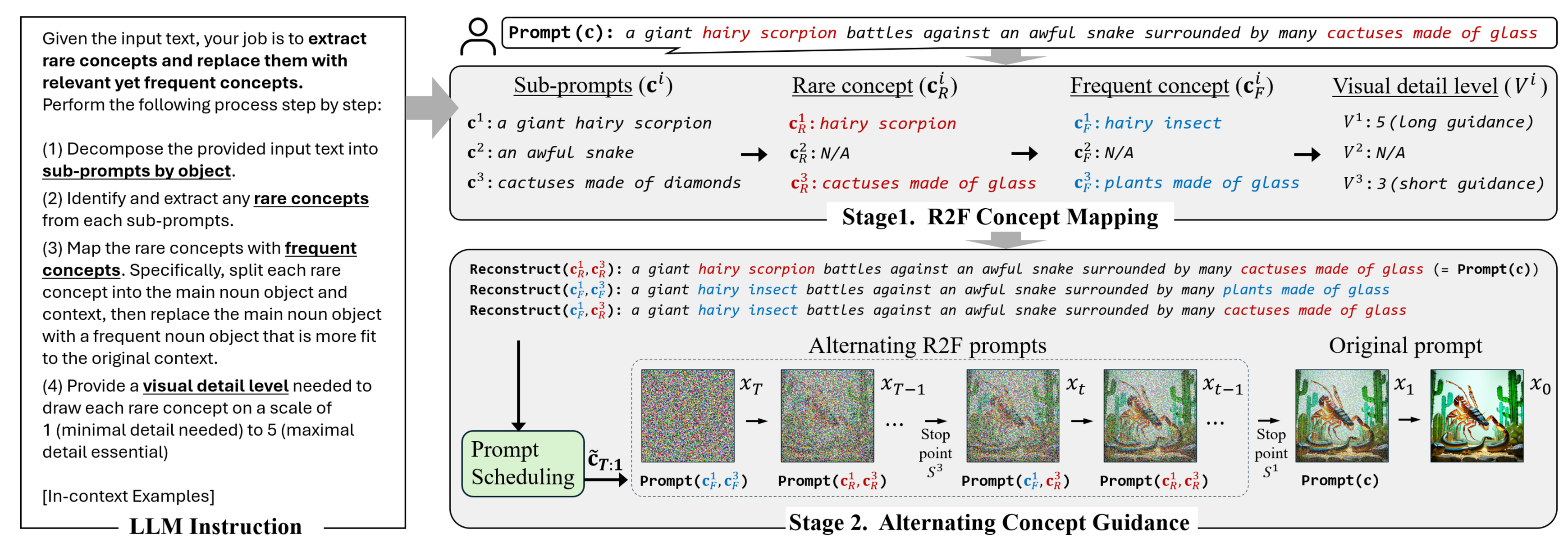


Rare-to-Frequent (R2F)

- **LLM finds** the frequent concept and **guides diffusion** sampling steps given a rare concept
- **Flexible** to arbitrary LLMs and diffusion backbones
- **Compatible** with region-guided methods → **R2F+ (See paper)**



Detailed Framework



1. Do **prompt decomposition** by objects
2. Finds **rare-to-frequent concept mapping** for each sub-prompt
3. Get **visual-detail-level** to draw each sub-concept
4. **Interpolate or alternate the rare and frequent concept** prompts throughout diffusion sampling steps until the stop point determined by the visual-detail-level

RareBench: A New T2I Benchmark

- Our proposed benchmark consisting of **diverse prompts with rare concept compositions**
- Generated by GPT and additionally inspected by human

rare object-attribute combinations

Datasets	RareBench	DVMP	T2I-CompBench
# Prompts	320	200	2400
% Rareness	98.1	62.0	17.4

Attributes used)

Property	hairy, horned, wooly, bearded, mustachioed, thorny, spiky, wrinkled, spotted, wigged, hairless
Shape	banana-shaped, star-shaped, ax-shaped, butterfly-shaped, oval-shaped, donut-shaped, hand-shaped, gear-shaped, heart-shaped, diamond-shaped
Texture	flower-patterned, zebra-striped, tiger-striped, black-white-checkered, made of marble, made of diamonds, made of plastic, made of glass, made of steel, made of cloud
Action	dancing, walking, running, crawling, flying, swimming, driving a car, yawning, smiling, crying, cheerleading

Results

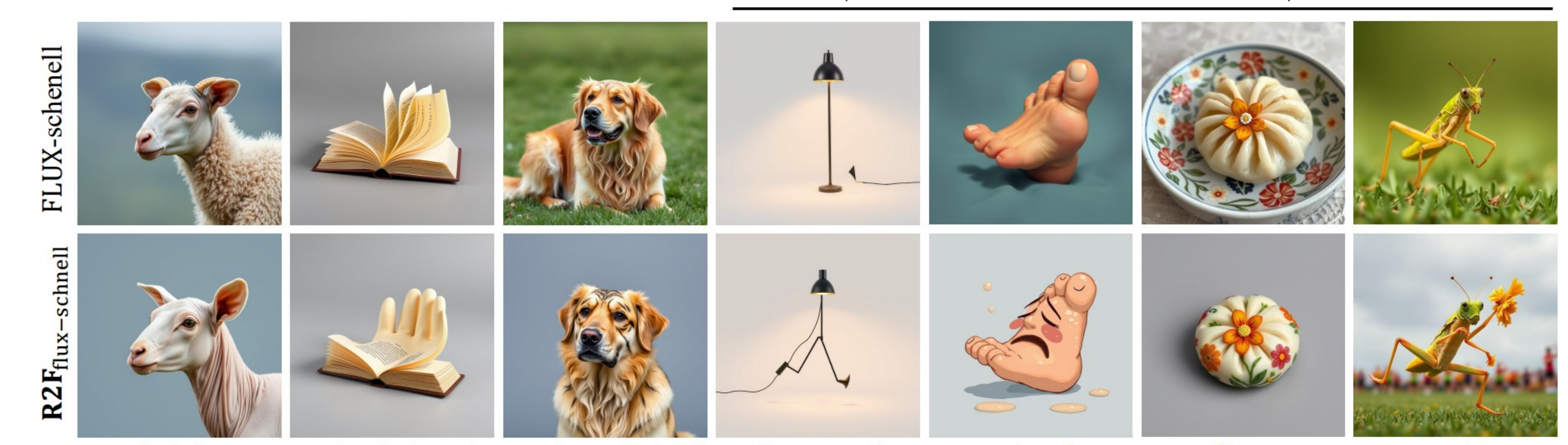
1. **Better T2I alignment** than SOTA pretrained & region-guided diffusions

Models	Single Object: Property GPT4	Property Human	Shape GPT4	Shape Human	Texture GPT4	Texture Human	Action GPT4	Action Human	Complex GPT4	Complex Human	Multi Objects: Concat GPT4	Concat Human	Relation GPT4	Relation Human	Complex GPT4	Complex Human
SD1.5	55.0	49.6	38.8	51.7	33.8	55.6	23.1	47.5	36.9	44.2	23.1	29.8	24.4	20.0	36.3	19.8
SDXL	60.0	55.2	56.9	57.7	71.3	63.3	47.5	59.0	58.1	60.4	39.4	35.8	35.0	28.8	47.5	41.7
PixArt	49.4	59.6	58.8	60.8	76.9	69.0	56.3	69.8	63.1	70.6	35.6	38.1	30.0	31.0	48.1	42.7
SD3.0	49.4	66.9	76.3	79.0	53.1	62.7	71.9	73.3	65.0	70.8	55.0	64.6	51.2	55.2	70.0	63.5
FLUX	58.1	63.8	71.9	70.0	47.5	61.7	52.5	67.1	60.0	67.3	55.0	57.3	48.1	50.6	70.3	66.7
SynGen	61.3	46.9	59.4	44.8	54.4	57.3	33.8	48.3	50.6	49.0	30.6	35.8	33.1	23.5	29.4	20.4
LMD	23.8	41.5	35.6	46.0	27.5	51.5	23.8	45.2	35.6	39.8	33.1	23.5	34.4	30.4	33.1	21.0
RPG	33.8	47.1	54.4	57.1	66.3	60.8	31.9	44.0	37.5	38.1	21.9	25.6	15.6	14.4	29.4	39.6
ELLA	31.3	49.6	61.6	54.8	64.4	61.9	43.1	53.8	66.3	60.6	42.5	45.6	50.6	39.6	51.9	47.9
R2F	89.4	86.3	79.4	80.6	81.9	71.5	80.0	79.4	72.5	75.6	70.0	71.3	58.8	57.9	73.8	67.3

2. **Flexible** to LLMs (LLaMA-3, GPT-4o) & Diffusions (SDXL, SD3, FLUX)

Models	Property	Shape	Texture	Action	Complex
SD3.0	49.4	76.3	53.1	71.9	65.0
$\mathrm{R2F}_{\mathrm{LLaMA3}}$	**81.9**	**77.1**	**76.3**	**78.8**	**67.7**
$\mathrm{R2F}_{\mathrm{GPT\text{-}4o}}$	**89.4**	**79.4**	**81.9**	**80.0**	**72.5**

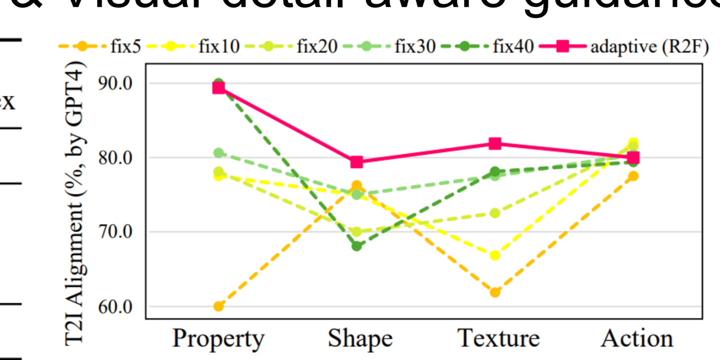
Models	Single Object: Property	Shape	Texture	Action	Complex	Multi Objects: Concat	Relation	Complex
IterComp	63.8	66.9	61.3	65.6	61.9	41.3	29.4	53.1
$\mathrm{R2F}_{\mathrm{itercomp}}$	**78.1**	**77.5**	**79.4**	**66.9**	**63.9**	**41.5**	**36.6**	**53.4**
SD3.0	49.4	76.3	53.1	71.9	65.0	55.0	51.2	70.0
$\mathrm{R2F}_{\mathrm{sd3.0}}$	**89.4**	**79.4**	**81.9**	**80.0**	**72.5**	**70.0**	**58.8**	**73.8**



A hairless sheep | A hand-shaped book | A tiger-striped golden retriever | A lamp running with legs | A crying foot | A flower-patterned dumpling | A grasshopper cheerleading

3. **Effective module design** (Interpolation & Visual-detail-aware guidance)

Models	Single Object: Property	Shape	Texture	Action	Complex	Multi Objects: Concat	Relation	Complex
SD3.0	49.4	76.3	53.1	71.9	65.0	55.0	51.2	70.0
Interpolate	85.5	77.0	69.4	74.6	71.7	54.0	53.8	71.3
Composable	82.5	76.3	58.1	68.1	67.5	63.1	51.9	61.9
P2P	71.3	46.3	46.9	38.8	52.5	31.3	32.5	33.8
R2F	**89.4**	**79.4**	**81.9**	**80.0**	**72.5**	**70.0**	**58.8**	**73.8**



4. **(R2F+) Better spatial composition** with attention-control integration

SD3.0 R2F+ SD3.0 R2F+ SD3.0 R2F+ SD3.0 R2F+

A hairless yak and a spotted tulip | A thorny building is overshadowing a bearded snowman | A teal spiky metal camera and a brown banana | An oblong cucumber and a teardrop plum

Takeaways

✓ **Exposing relevant-yet-frequent concepts** improves **compositionality** of diffusion models

✓ **R2F** is a **scalable framework** that leverages **LLMs to identify rare concepts in any text** and provide guidance for their generation